Auditory Spectral Integration Effects in Dynamic Consonant-Vowel /da/-/ga/ F3 Transitions

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation with distinction in Speech and Hearing Sciences in the undergraduate colleges of The Ohio State University

By

Lisa A. Wackler

The Ohio State University

June 2007

Project Advisors: Dr. Robert Fox and Dr. Ewa Jacewicz

ABSTRACT

In speech perception, identification of place of articulation is often dependent upon formant transitions (frequency changes in vocal resonances). A number of experiments have shown that the alveolar and velar place distinction in initial stop consonants (/d/ and /g/) can be cued by the slope of the third formant (F3) transition. Previous studies have established that these place distinctions can be made with both actual and "virtual" transitions. However, the question remains as to the place at which this auditory processing takes place: Does it occur in the auditory periphery or as a function of central auditory processing? The experiment presented here addressed the extent to which the perception of synthetic /da/-/ga/ syllables could be cued by virtual transitions in both diotic and dichotic conditions. Both virtual and actual F3 transitions produced similar identification functions but the efficacy of the virtual transition was somewhat reduced in the dichotic condition.

TABLE OF CONTENTS

I. INTRODUCTION AND LITERATURE REVIEW

The ability to produce and understand speech is often taken for granted and little thought is given to the underlying complex cognitive processes that allow communication to occur. Our ability to share experiences, transmit knowledge, and exchange thoughts and ideas has unquestionably been largely dependent upon speech throughout the development of human culture. Speech has proven to be the most convenient form of communication for human society at large, and a great deal of research has sought to gain a more thorough understanding of how speech is produced, perceived, and understood.

Production of Speech

The production of the individual sounds that comprise speech is an intricate combination of specialized movements of vocal organs, beginning with the respiratory system pushing air out of the lungs. In nearly all speech sounds, the basic source of power is the lungs, which send an air stream flowing through the trachea and the larynx, and into the pharynx and the oral cavity. Upon entering the larynx, the airstream passes between the two small muscular folds, called the vocal folds. If the vocal folds are adjusted so that there is only a comparatively narrow passage between them, the airstream sets the vocal folds into rapid vibration, producing their characteristic buzz. In general, sounds produced when the vocal folds are apart are referred to as voiceless sounds, while those produced when the vocal folds are vibrating are referred to as voiced sounds.

The acoustic energy that emerges from the vocal folds is modified by the acoustic properties of the vocal tract, which depend primarily on its size and shape. While the size of the vocal tract remains constant within each speaker, the shape of the vocal tract (consisting of the pharynx, mouth, and nose) can be varied extensively by moving the soft palate, tongue, lips and jaw; referred to as the articulators. This process of adjusting the vocal tract shape to produce distinctive, varying speech sounds is called articulation. During speech, the movement of these articulators continually alters the shape of the vocal tract and its acoustic characteristics, which enables us to produce the different sounds of speech. Combining these movements with the dynamic action of the vocal folds produces a wide variety of consonant sounds, which are typically classified according to three major dimensions: place of articulation, manner of articulation, and voicing.

The action of the vocal folds is responsible for converting the energy of the airstream into acoustic energy, before it is subjected to the various modifications of the articulators, as it progresses through the vocal tract. The vocal folds' frequency of vibration is known as the fundamental frequency (F_o), and is the lowest frequency component present in a spectrum of speech sounds. Fundamental frequency is a factor of the tension and length of the vocal folds, as well as the air pressure from the lungs, which are continually being modified during speech. As the acoustic energy progresses toward the lips, the vocal tract acts as a resonator, meaning that it has certain natural frequencies of vibration, or resonant frequencies. The vocal tract responds more readily to a component of the vocal folds' vibration whose frequency is similar to its own resonant

frequency, than to a sound wave of another frequency. These components will be emphasized and the spectrum of the sound emerging from the lips will "peak" at the resonant frequencies of the vocal tract (Denes and Pinson, 1993). The spectrum of the speech wave will have a peak at a number of different frequencies because the vocal tract emphasizes the harmonics of the vocal fold wave at each of the vocal tract's resonant frequencies. Since the values of the resonant frequencies are determined by the shape and size of the vocal tract, there will be spectral peaks at different frequencies as the shape of the vocal tract is changed.

These resonances of the vocal tract are called formants, and they are particularly important in our understanding of speech perception and the ability of the human auditory system to distinguish between various speech sounds. Formant frequencies depend greatly on whether and where the articulatory movements of the tongue and lips obstruct the oral cavity; every configuration of the vocal tract has its own set of characteristic formant frequencies. These characteristics of the acoustic signal are important in providing the information necessary to distinguish between the various places and manners of articulation.

Examination of the acoustic characteristics of speech waves has proven successful in producing a great deal of information about formants. While we are able to effectively describe the ways in which the vocal tract and oral cavity move in order to produce the various sounds of speech, there is still much to be learned about the complex mechanisms involved in the perception of speech. Of interest is how the human listener processes these complex sounds. There has been a

great amount of research attempting to answer the question of what the human ear pays attention to in order to accurately distinguish between different speech sounds and, further, to successfully understand speech.

Formant Averaging: the Center-of-Gravity Effect

Early research investigating the importance of formants in speech perception was conducted by Delattre *et al.* (1952), addressing the role formant frequencies play in vowel perception. This study showed that the phonetic quality of back vowels synthesized with only the first two formants could be matched to a vowel containing only a single formant. The success of the match proved to be dependent upon the relationship between the frequencies and amplitudes of the two close formants and the peak frequency of the single formant vowel. Variation in the relative intensity ratio between the two close formants in these synthetic back vowels produced a change in the frequency of the single formant to which it was best matched. More specifically, as the ratio of the level of F2 to F1 was increased, a significant systematic raising occurred for the center frequency of the single-formant vowel whose quality was being matched. It was proposed that an auditory mechanism must effectively average two formants that are relatively close in frequency in order to achieve this effect. This phenomenon, known today as formant averaging, spectral integration, or the center-of-gravity (COG) effect, suggests that the auditory system performs an additional filtering of vowels beyond the cochlea.

A further exploration of this COG effect led to several studies investigating the integration of formants during vowel identification in vowel matching experiments. A number of experiments conducted by Chistovich and colleagues (Bedrov *et al.*, 1978; Chistovich and Lublinskaja, 1979; Chistovich *et al.*, 1979) showed that the predictable shift in the matching frequency of the single formant occurred when the two close formants fall within a "critical distance" or "critical formant separation" of about 3.5 bark. What is referred to as one bark is an empirically-derived common value that is thought to represent the width of one of the presumed bank of adjacent filters used to model basilar membrane mechanics (Scharf, 1972). This 3.5 bark critical distance is interpreted by some to denote the spectral resolving power of the auditory periphery. It was proposed that within this critical distance the changes to the relative amplitude ratios between the two formants changed their combined spectral center of gravity (COG) and it was to this spectral COG that the frequency of the single formant was being matched. This work showed that when two vowel formant peaks are separated by less than 3.5 bark, they are perceptually integrated into a single perceived peak ("perceptual formant"), with a frequency that is closer to that of the stronger formant (Chistovich *et al.*, 1979). The amplitudes of the two formants play an important role insofar as a change in their ratio is equivalent to a *frequency* change of a single-formant vowel which approximates their quality (Chistovich and Lublinskaja, 1979). The 3.5-bark critical distance indicates a possible limit on spectral integration in that the COG effect disappears with larger formant separation.

Psychoacoustic studies of Virtual Pitch

Insights into this type of auditory filtering can be drawn from early psychoacoustic work by Feth (1974) and Feth and O'Malley (1977). This research investigated the spectral pitch of complex tones using two-component complex tone pairs that had identical envelopes but differed in fine structure (Voelcker 1996a; b). The results were consistent with the COG hypothesis of Chistovich *et al.*, showing decreasing discriminability as the separation of the two components increased to 3.5 bark. The similarity of the two-tone resolution results of Feth and O'Malley (1977) to the critical distance observed in vowel matching tasks serves to further demonstrate the possibility of a common mechanism, i.e., an auditory spectral resolving power. Recently, Xu *et al.* (2004) confirmed both the results of Chistovich and Lublinskaja (1979) and Feth and O'Malley (1977) with a set of American English listeners responding to the same two types of signals (two-formant synthetic vowels and complex two-tones). In their study, Xu *et al.* concluded that both the complex-tone discriminability and the spectral integration limits reflect the same auditory spectral resolving power, and therefore further suggests that the auditory processing of complex auditory signals at the intermediate stage is the same for speech and nonspeech signals.

Spectrally Dynamic COG Effects

An important limitation of the past research was that the signals used were spectrally static; that is the parameters of the signals remain constant for the entire duration of the sound. These stationary sounds are unnatural, as they normally do not occur in speech. Real speech signals are typically dynamic, with

formant frequencies and amplitudes changing, sometimes rapidly, over time. It became evident that since dynamic signals are more prevalent in the natural speech environment, they would be more appropriate for studying integration effects. In a landmark study, Lublinskaja (1996) used two Russian diphthongal vowels to investigate the auditory system's ability to attend to a dynamic spectral COG. The ratios of the amplitudes of two relatively closely spaced formants were modified over time to attempt to produce the perception of a non-stationary, diphthongal vowel. Lublinskaja reported that these virtual formant changes did in fact produce a diphthongal-vowel percept when the critical distance between the modified formants (F2 and F3) was less than 3.5 bark. However, when the distance between F2 and F3 was greater than the critical distance of 3.5 bark the percept was that of a stationary vowel. In addition, moving the frequency separation from 3.5 to 4.5 bark produced systematically more stationary percepts as the size of the separation increased. Accordingly, it can be concluded that the spectral integration limit for this type of dynamic vowel is between 3.5 and 4.5 bark, which is somewhat larger than the value determined for static vowels.

Building on this research, Xu *et al.* (2004) extended the investigation of the COG effect observed in diphthongal vowels to consonant-vowel (CV) transitions. The study sought to analyze whether virtual frequency (VF) and frequency modulated (FM) transitions are processed in a way that makes them perceptually equivalent to that of a synthetic formant transition, looking specifically at the CV transitions in /da/ and /ga/. The results showed no significant difference in the identification responses between stimulus types, revealing that the dynamic

change caused by amplitude modulation is a phenomenon comparable to a frequency change. Xu *et al.* concluded that this "virtual" frequency change is processed similarly in speech and nonspeech signals and appears to occur more centrally in the auditory system.

Also of interest is a study by Fox *et al.* (2006) which assessed the extent to which the perception of synthetic /da/-/ga/ and /ta/-/ka/ syllables could be cued by virtual formant transitions or virtual bursts brought about by spectral COG effects. The results showed that listeners were able to accurately identify consonants along a /da/-/ga/ continuum using these stimulated, virtual formant transitions, and further that the dynamic cues to place of articulation in these syllables were perceived and interpreted in a near equal fashion by the auditory system regardless of the transition type.

One important unanswered question is whether auditory spectral integration occurs in the auditory periphery (within the cochlea prior to central auditory processing), or within higher auditory processing centers as a part of central auditory processing. In order to answer this question and further investigate the COG phenomenon, the present study attempted to determine the extent to which spectral integration of auditory signals corresponding to the third formant transition (which produces a /d/-/g/ contrast) occurs. The existence of dynamic cues to place of articulation, found in synthesized transitions along an actual /da/-/ga/ continuum, as well as along a /da/-/ga/ continuum using simulated, virtual formant transitions, was further investigated; extending the investigation

to dichotic listening paradigms. Of particular interest is the contrast between diotic presentation (identical auditory stimuli presented to both ears) and dichotic presentation (presentation of different parts of the stimuli, simultaneously, to each ear) of the stimuli. The experiment was designed with both diotic and dichotic presentations of stimuli in an attempt to substantiate the claims that central mediation is required to arrive at a unified percept. If the diotic and dichotic presentation conditions produce similar identification functions, this would argue that the human auditory system is able to integrate different sounds from both ears to arrive at a unified percept, suggesting that this type of processing is happening at central levels of the auditory system. As with the previous studies investigating dynamic consonant-vowel transitions, selecting /da/ and /ga/ as CV units for the present investigation is particularly appropriate because it has been well demonstrated that the direction of the F3 transition can effectively cue the place of articulation (Mann and Liberman, 1983; Whalen and Liberman, 1987; Fox *et al.*, 1997). The direction of the F3 transition is responsible for the auditory distinction between /da/ and /ga/: a rising transition leads to the perception of /g/ and a falling transition specifies /d/.

Therefore, since it is this formant transition characteristic of the acoustic signal that provides the information necessary to distinguish between various places and manners of articulation, a more thorough understanding of how the auditory system processes this information will provide new insight into the peripheral and central mechanisms underlying human speech perception.

II. METHODS

A. Stimuli

The stimuli were designed so that there were six different /da/-/ga/ series sets whose steps differed in terms of two variables; the method of presentation of the tokens (diotic or dichotic) and the type of F3 transition within each token (actual or virtual F3 transition).

Diotic, Actual F3 Transition Series

These stimuli were created using the parallel branch of the Klatt synthesizer using the .kld option in HLSYN (Sensimetrics, 1997) with a sampling rate of 11025 Hz. Each token had a duration of 250-msec, consisting of a 50-msec transition portion (corresponding to the consonant-vowel formant transitions) and a 200-msec steady-state vowel portion. The 50-msec transition portion of F1 and F2, as well as the 200-msec steady-state portion for all three formants (F1, F2, and F3) were identical in each token in the series. Over the first 50 msec, F1 increased from 443 Hz to 700 Hz and F2 decreased from 1520 Hz to 1220 Hz. The frequencies of F1 and F2 then remained unchanged at 700 Hz and 1220 Hz, respectively, over the final 200 msec steady-state vowel portion of the token. The frequency of F3 for the steady-state portion of the vowel was a constant 2600 Hz. The tokens differed in terms of the 50-msec transition portion of F3. These nine F3

transitions had different onset frequencies that were equally spaced, ranging from 1800 Hz to 2600 Hz, each with an offset frequency of 2600 Hz. The onset and offset frequencies of these nine F3 transitions are shown in Table 1. In addition, Figure 1 is a schematic representation of the diotic, actual F3 transition series.

Series Step	Transition Onest	Transition Offest
1	1800	2600
2	1900	2600
3	2000	2600
4	2100	2600
5	2200	2600
6	2300	2600
7	2400	2600
8	2500	2600
9	2600	2600

Table 1. Onset and Offset frequencies of Actual F3 Transitions

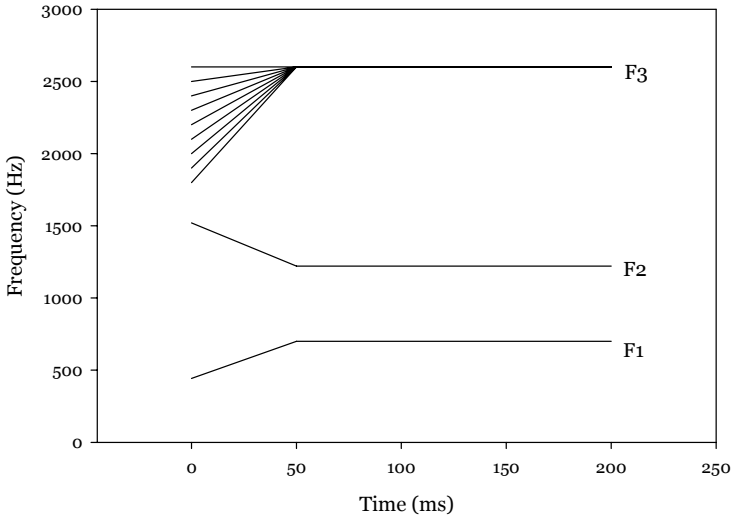

Figure 1. Schematic Representation of the diotic, actual F3 transition series; note that the frequency of the F3 onset varies from 1800 Hz to 2600 Hz, as outlined in Table 1.

Diotic, Virtual F3 Transition Series

The construction of these stimuli consisted of two separate processes, involving creating a virtual F3 transition and inserting it into the first 50 msec of a base consonant-vowel token. The base token was created using the parallel branch of the Klatt synthesizer, again using the HLSYN with the .kld option and a sampling rate of 11025 Hz. The base token consisted of a 50-msec transition portion and a 200-msec steady-state vowel portion. The transition portion of the base consisted of the F1 and F2 transitions only, whereas the steady-state portion contained the first three steady-state formants of the vowels. The 50-msec transition portion of F1 and F2, as well as the 200-msec steady-state portion for all three formants (F1, F2, and F3) were again identical in each token in the series. Over the first 50 msec, F1 increased from 443 Hz to 700 Hz and F2 decreased from 1520 Hz to

1220 Hz. The frequencies of F1 and F2 then remained unchanged over the final 200 msec steady-state vowel portion of the token. The frequency of F3 for the steady-state portion of the vowel was a constant 2600 Hz. The tokens differed in terms of the 50-msec transition portion of F3.

The virtual F3 transitions were created by changing the relative intensities (intensity ratios) of two 50-msec sinewave pairs. The frequencies of the sinewaves were multiples of the fundamental frequency (F_0=120 Hz), and were just above and just below the frequencies where the actual F3 transitions occurred (see Table 2). These sinewave pairs were created using the tone generator option in Adobe Audition, and were created with equal amplitudes. The relative intensities of these pairs of sinewaves were then adjusted in nine different steps, and then combined, so that the movement of the spectral center-of-gravity of the tokens changed in the same fashion as the frequencies of the F3 transition in the tokens with an actual F3 transition. The relative intensities of the onsets and offsets are shown in Table 2. The spectral center-of-gravity at each temporal location in the F3 transition was based on the intensity ratio between the lower and higher sinewave pairs. The overall mean rms of the combined pairs of sinewaves was adjusted to within 1dB of the average rms value of the actual F3 transitions.

Series Step	Relative Intensity Onest (P1)	Relative Intensity Offest (P1)	Relative Intensity Onest (P2)	Relative Intensity Offest (P2)
1	93.75%	10.42%	6.25%	89.58%
2	83.33%	10.42%	16.67%	89.58%
3	72.92%	10.42%	27.08%	89.58%
4	62.50%	10.42%	37.50%	89.58%
5	52.08%	10.42%	47.92%	89.58%
6	41.67%	10.42%	58.33%	89.58%
7	31.25%	10.42%	68.75%	89.58%
8	20.83%	10.42%	79.17%	89.58%
9	10.42%	10.42%	89.58%	89.58%

Table 2. Relative intensities of onsets and offsets of sinewave pairs. (Lower pair, harmonics used = 1680 Hz and 1800 Hz, mean = 1740 Hz; Higher pair, harmonics used = 2640 Hz and 2760 Hz, mean = 2700 Hz).

The 50-msec virtual transitions were then inserted into a copy of the base token using Audition, with the onset of the transitions synchronized with the onset of the F1 and F2 formants in the base token. Figure 2 is a schematic representation of the final three-formant tokens.

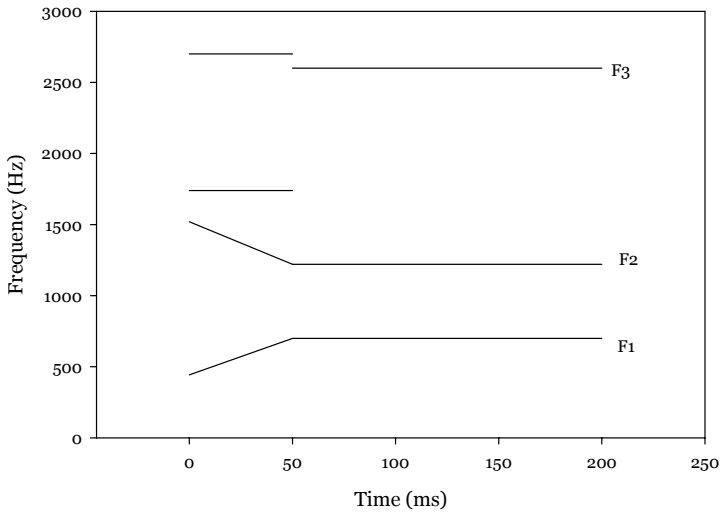

Figure 2. Schematic representation of the diotic, virtual F3 transition series; note that the center frequencies of the two 50-msec sinewave pairs remain constant, with the variation lying in the modified relative intensities, as outlined in Table 2.

Dichotic, Actual F3 Transition Series

In order to present the stimuli dichotically it was necessary to create two separate channels, which could be used to simultaneously present different parts of the token to each ear. These stimuli were constructed with a base token (containing F1 and F2 with the same frequencies and durations previously used) in one channel, and the nine different F3 steps (containing both the 50-msec transition portion and the 200-msec steady-state vowel portion) in the other channel. Figure 3 is a schematic representation of the dichotic, actual F3 transition series. The stimuli were created using the parallel branch of the Klatt synthesizer using the .kld option in HLSYN with a sampling rate of 11025 Hz, and the F3 portion was then inserted into a copy of the base

token using Audition, with the onset of the F3 formant synchronized with the onset of the F1 and F2 formants in the base token. The nine F3 formant transitions were identical to those created in the diotic files, with different onset frequencies that were equally spaced, ranging from 1800 Hz to 2600 Hz. Each of the nine F3 transitions had an offset frequency of 2600 Hz, and the frequency of F3 for the steady-state portion of the vowel was a constant 2600 Hz.

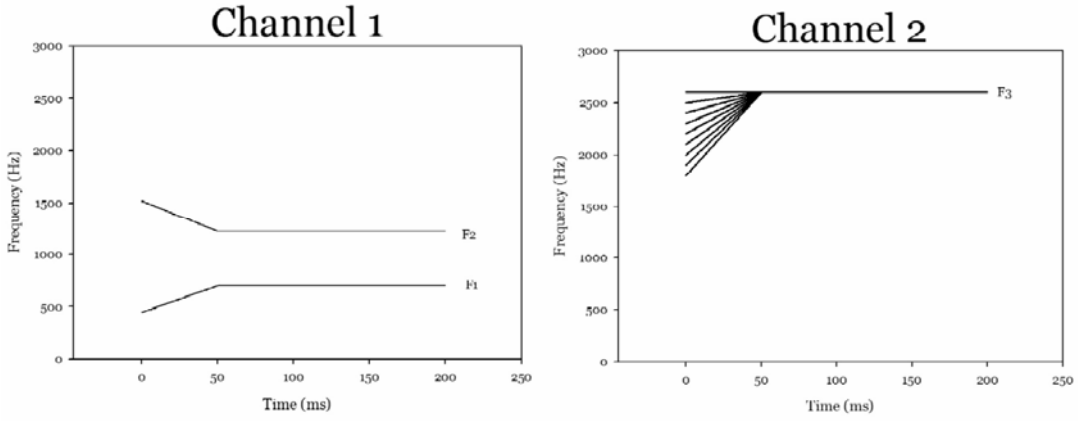

Figure 3. Schematic Representation of the dichotic, actual F3 transition series. Channels 1 and 2 were presented simultaneously to separate ears.

Dichotic, Virtual F3 Transition Series

The dichotic virtual stimuli were created, similarly to the dichotic stimuli with an actual F3 transition, with separate channels for the presentation of separate pieces of the token to each ear. There were three dichotic virtual series versions, differing by the ways in which the stimulus was divided between the ears.

Version one was constructed with the base token in one channel and the F3 steps and steady-state portion in the other channel. The base token contained F1 and F2. Over the first 50 msec, F1 increased from 443 Hz to 700 Hz and F2 decreased from 1520 Hz to 1220 Hz. The frequencies of F1 and F2 then remained unchanged over the final 200 msec steady-state vowel portion of the token. The F3 portion of each stimulus consisted of a 50-msec virtual transition, created in the same manner as those previously described, with the relative intensities of two sinewave pairs being manipulated in order to change the spectral center-of-gravity in the same fashion as the frequencies of the F3 transition in the tokens with an actual F3 transition. Again, the final 200-msec of the F3 remained constant at 2600 Hz. Figure 4 is a schematic representation of version one of the dichotic, virtual F3 transition series.

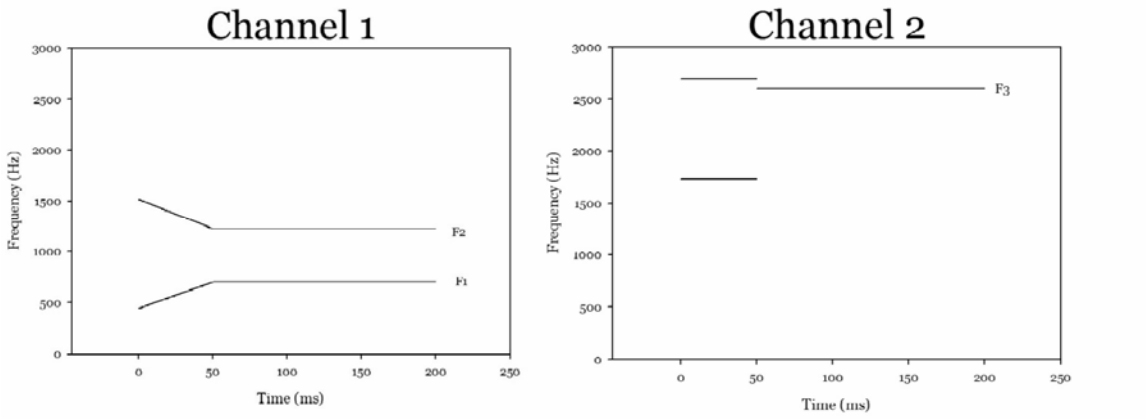

Figure 4. Schematic representation of the diotic, virtual F3 transition series, **version 1**; note that the the two 50-msec sinewave pairs remain constant, and are in the same channel. Channels 1 and 2 were presented simultaneously to separate ears.

Versions two and three were different from version one in that the two sinewave pairs necessary to create a virtual F3 transition were split between the two channels. Version two had the entire 250-msec F1 and F2 in one channel, as well as the lower frequency (mean = 1740 Hz) 50-msec sinewave pair, while the other channel contained the higher (mean = 2700 Hz) frequency 50-msec sinewave pair and the 200-msec steady state portion of F3. Version three was constructed in a similar manner, the key difference being that the sinewave pairs were split so that they were opposite the arrangement used in version two. The entire 250-msec F1 and F2 was in one channel along with the higher frequency (mean = 2700 Hz) 50-msec sinewave pair, and the lower frequency (mean = 1740 Hz) 50-msec sinewave pair and the 200-msec steady state portion of F3 were in the other channel. Figures 5 and 6 are schematic representations of versions 2 and 3, respectively, of the dichotic, virtual F3 transition series.

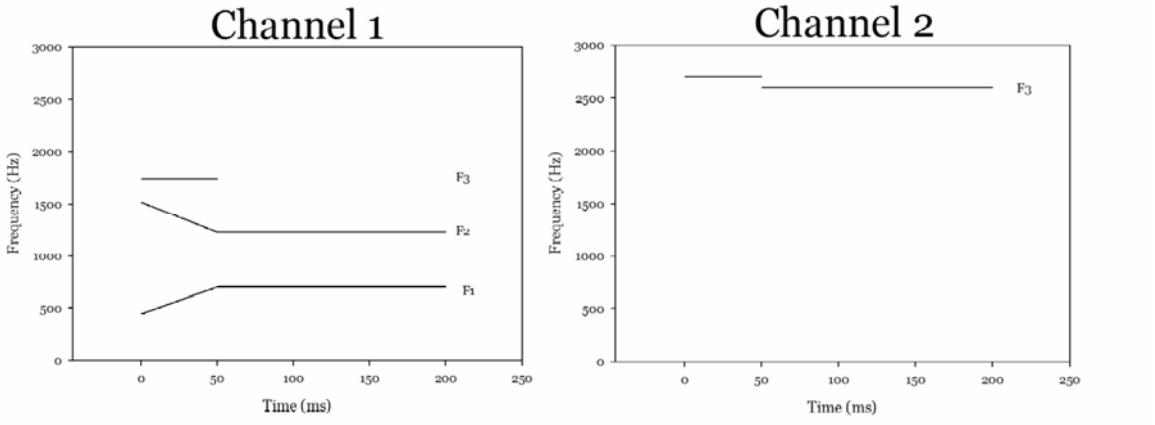

Figure 5. Schematic representation of the diotic, virtual F3 transition series, **version 2**; note that the the two 50-msec sinewave pairs remain constant, and have been split between the two channels. Channels 1 and 2 were presented simultaneously to separate ears. (F3 Sinewave mean frequencies = Channel 1: 1740 Hz, Channel 2: 2700 Hz.)

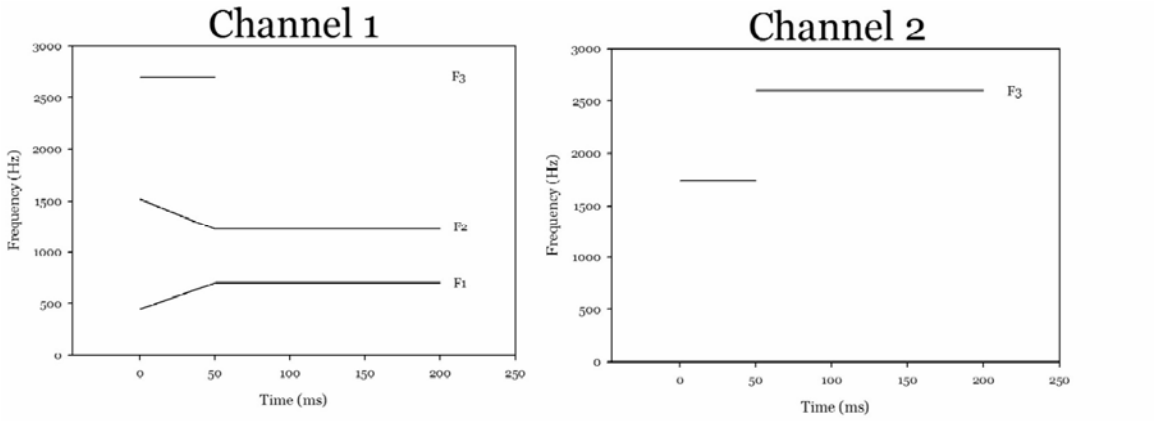

Figure 6. Schematic representation of the diotic, virtual F3 transition series, **version 3**; note that the the two 50-msec sinewave pairs remain constant, and have been split between the two channels (opposite version 2). Channels 1 and 2 were presented simultaneously to separate ears. (F3 Sinewave mean frequencies = Channel 1: 2700 Hz, Channel 2: 1740 Hz.)

B. Listeners

Fourteen listeners (five men and nine women) with no known history of hearing impairment participated in the experiment. The listeners ranged in age from 19 to 28 years, and were paid volunteers. All subjects were native speakers of American English. They were paid $16.00 ($8.00/hour, for two one-hour sessions) for their participation. Two additional listeners were unable to complete the listening task; their results were not included in the analyses.

C. Procedure

All signals were presented with TDH-49 headphones at a comfortable listening level to a subject seated in a sound-attenuating booth. A single-interval two-alternative forced choice identification task was used with the response choices "*da as in 'dot'*" and "*ga as in 'got'*" displayed on two separate halves of a computer monitor. Subjects were asked to indicate whether they heard a *da* or a *ga* for each token presented by clicking a mouse on the appropriate section of the display. There were 135 stimuli presented randomly in each of six sets (9 tokens x 15 repetitions) blocked by the token type. Each subject participated in two separate one-hour sessions, with three sets of 135 stimuli presented in each session. The presentation order of sets was counterbalanced across listeners. The sets were:

1) Diotic, actual transition
2) Diotic, virtual transition
3) Dichotic, actual transition
4) Dichotic, virtual transition, version 1
5) Dichotic, virtual transition, version 2
6) Dichotic, virtual transition, version 3

In order to familiarize the subject with the stimulus set and the task, prior to the listening task for each blocked set of stimuli the subject heard each endpoint stimulus for that stimulus series (step 1 and step 9) eight times.

As each stimulus token was played, the appropriate response area turned red on the computer display. Then, the subject was presented with a nine-item randomized self-check, instructed to guess the right answer, and the appropriate response again turned red on the computer display. Following the presentation of these examples, the subject had a fifteen-item practice, with no feedback. After the practice was completed, the experimenter answered any questions the subject had. The subject was then presented with the endpoint examples and the randomized self-check again before the actual listening task began. Each of six sets lasted approximately twenty minutes.

III. RESULTS AND DISCUSSION

Diotic Condition

Shown in Figure 7 are the identification functions of the responses to the actual F3 transition series and the virtual F3 transition series, both presented diotically. The two identification functions have a similar shape, although it is clear that the slope of the identification function for the virtual F3 transition series is more gradual than that of the actual F3 transition series.

The locations of the /da/-/ga/ category boundary (the 50% cross-over point) along the F3 onset axis for each individual step were calculated using a one-way between-subject Probit Analysis with the F3 transition type (actual or virtual) as the factor. This test showed that there was a significant difference in the mean category boundary ($F(1,12)=13.183$, $p=0.003$, $\eta^2=0.523$) between the actual F3 transition series and the virtual F3 transition series. Listeners provided slightly more *da* responses to the virtual F3 transition series (mean category boundary=2121 Hz) than to the actual F3 transition series (mean category boundary=2229 Hz). The results indicate that listeners demonstrated some bias toward *da* in the virtual F3 transition series, while no bias was demonstrated in the actual F3 transition series.

Next, the number of *da* responses were analyzed using a two-way within-subject analysis of variance with the factors F3 transition type (actual or virtual) and F3 onset frequency (series step). There was a significant main effect of transition

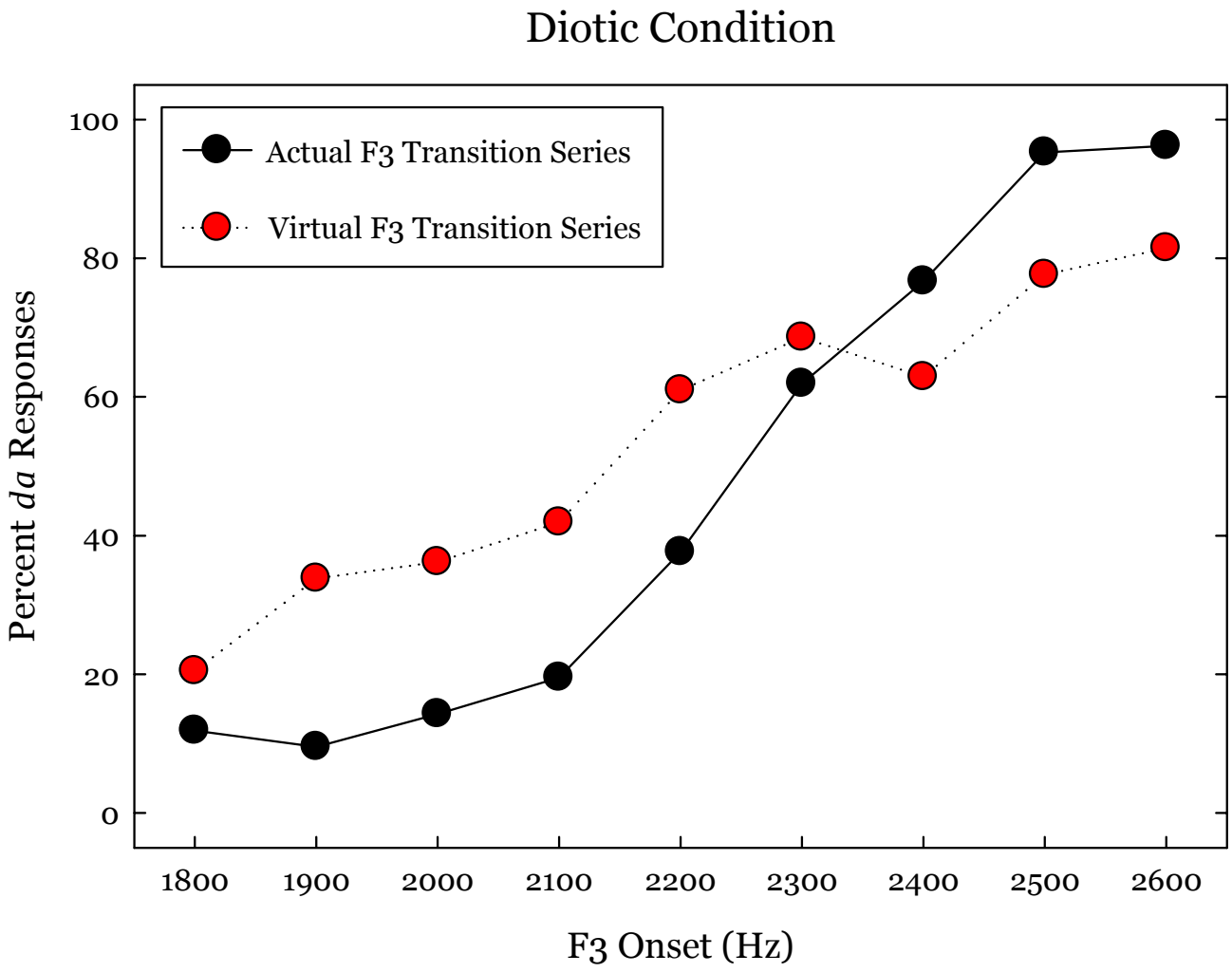

Figure 7. Identification functions of responses to the actual F3 transition series and the virtual F3 transition series, presented diotically. Values along the abscissa represent the onset frequency used in the synthesis of the actual formant transition or, for the virtual transitions, the effective onset frequency of the spectral center-of-gravity for the two intensity-modulated sinewave pairs.

type ($F(1,13)=5.250$, $p=0.039$, $\eta^2=.288$), and, as expected, a significant main effect of F3 onset frequency ($F(8,104)=115.88$, $p<0.001$, $\eta^2=0.899$). There was also a significant F3 transition type by F3 onset frequency interaction ($F(8,104)=16.45$, $p<0.001$, $\eta^2=0.559$). This interaction stems from the fact that the slope of the identification function is somewhat more shallow in the virtual F3 transition series than in the actual F3 transition series.

In comparing the slopes of the two identification functions, a paired-samples t-test showed that there was a significant difference in the percent *da* responses given to the two different series ($t(13)=5.55$, $p<0.001$). The slope of the actual F3 transition identification function (mean=0.127) was greater then that of the virtual f3 transition (mean=0.076). It is clear that the actual F3 formant transitions are more salient than the virtual F3 formant transitions in the diotic condition. These results support the claim that the virtual F3 transitions do not provide as strong a cue for the /da/-/ga/ consonant distinction as do the actual F3 formant transitions.

Dichotic Condition

Shown in Figure 8 are the identification functions of the responses to the actual F3 transition series and the virtual F3 transition series, both presented dichotically. The two identification functions approximate the same general shape, however it is clear that the slope of the identification function for the virtual F3 transition series is much more gradual than that of the identification function for the actual F3 transition series. The slope of the identification

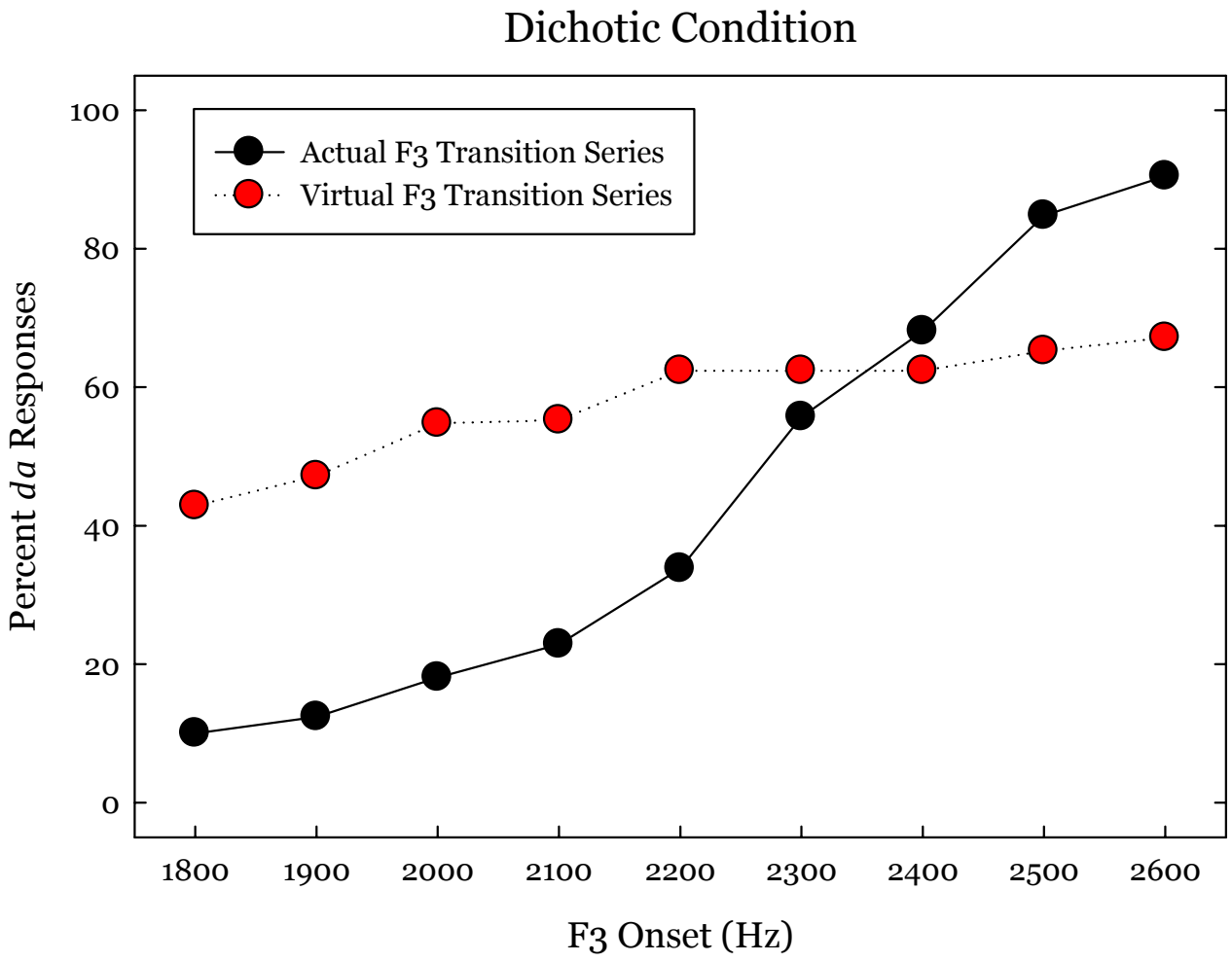

Figure 8. Identification functions of responses to the actual F3 transition series and the virtual F3 transition series, presented dichotically. Values along the abscissa represent the onset frequency used in the synthesis of the actual formant transition or, for the virtual transitions, the effective onset frequency of the spectral center-of-gravity for the two intensity-modulated sinewave pairs.

function of the actual F3 transition series shows a significantly more abrupt shift from primarily *ga* responses to primarily *da* responses, indicating that the actual F3 transitions are more salient in cueing the /da/-/ga/ distinction.

The locations of the /da/-/ga/ category boundary (the 50% cross-over point) along the F3 onset axis for each individual step were again calculated using a one-way between-subject Probit Analysis with the F3 transition type (actual or virtual) as the factor. This test showed that there was no significant difference in the mean category boundary ($F(1,9)=0.326$, $p=0.582$) between the actual F3 transition series and the virtual F3 transition series (actual F3 transition mean=2273 Hz, virtual F3 transition mean=2206 Hz).

Next, the number of *da* responses were analyzed using a two-way within-subject analysis of variance with the factors F3 transition type (actual or virtual) and F3 onset frequency (series step). There was a significant main effect of transition type ($F(1,13)=23.160$, $p<0.001$, $\eta^2=0.640$) and, as expected, a significant main effect of F3 onset frequency ($F(8,104)=47.082$, $p<0.001$, $\eta^2=0.784$). There was also a significant F3 transition type by F3 onset frequency interaction ($F(8,104)=20.85$, $p<0.001$, $\eta^2=0.616$). This interaction stems from the fact that the slope of the identification function is significantly more shallow for the virtual F3 transition series than in the actual F3 transition series.

In comparing the slopes of the two identification functions, a paired-samples t-test showed that there was a significant difference in the percent *da* responses

given to the two different dichotic series ($t(12)=8.183$, $p<0.001$). The mean slopes of the identification functions (actual F3 transition series=0.116, virtual F3 transition series=0.033) further illustrate this difference. It is clear that the actual F3 formant transitions are more salient than the virtual F3 formant transitions in the dichotic condition, as well, again showing that the actual F3 formant transitions are better able to cue the /da/-/ga/ consonant distinction.

A Closer Look at the Dichotic Virtual Conditions

Shown in Figure 9 are the identification functions of responses to the three versions of the dichotic virtual transition series. One can notice that while version one has a positive slope, versions two and three appear to be much more flat. As previously described, the three versions of the dichotic stimuli with virtual F3 formant transitions differ in terms of which pieces of the acoustic signal are being presented to each ear. While version one has the entire third formant in one ear, in versions two and three the sine wave pairs that comprise the virtual F3 transition are oppositely split between ears. The results show that listeners were able to more accurately identify the virtual transition, and therefore the /da/-/ga/ consonant distinction, in version one, where both sine wave pairs were presented to the same ear. It seems, in the cases of versions two and three, as though the auditory system could not effectively integrate the sine wave pairs when they were presented to separate ears. This suggests that the integration of these pieces of the signal may be occurring at a more peripheral level.

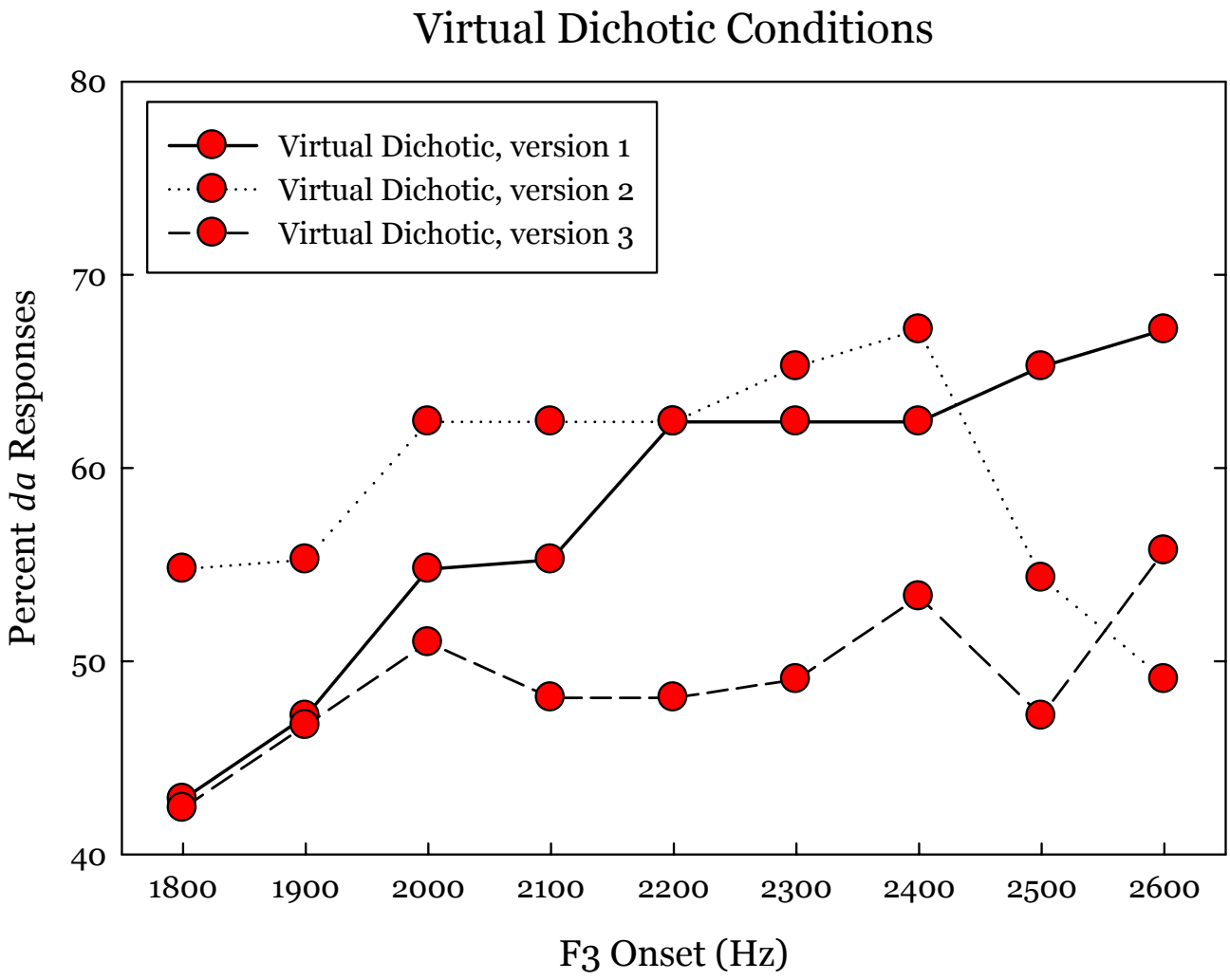

Figure 9. Identification functions of responses to the three versions of the dichotic, virtual transition series. Values along the abscissa represent the effective onset frequencies of the spectral center-of-gravity.

Summary and General Discussion

The experiment presented here addressed the extent to which the perception of synthetic /da/-/ga/ syllables could be cued by virtual transitions brought about by spectral COG effects, in both diotic and dichotic conditions. The results show that both actual and virtual dynamic cues to place of articulation in these syllables are perceived and interpreted in a similar fashion by the auditory system, although it is noteworthy that the virtual F3 formant transitions were not as salient as the actual F3 formant transitions in cueing the /da/-/ga/ distinction. In addition, listeners were better able to correctly identify the dynamic /da/-/ga/ consonant transition in the diotic condition than in the dichotic condition. It can be concluded that the stimuli with the actual F3 transitions were much more salient in cueing this consonant distinction, in both the diotic and the dichotic conditions, although the difference between the actual F3 transition and the virtual F3 transition was much more significant in the dichotic condition.

It is interesting to consider the process known as auditory streaming, or more specifically auditory stream segregation, whereby different sound elements are thought to be separated into different auditory objects in the auditory system's analysis of complex sounds. This process provides a potential explanation for why the dichotic condition was less salient than the diotic condition; it is possible that when the dichotic stimuli were played repeatedly for the listeners, their auditory systems separated the different components and processed them as separate streams, potentially lessening the effects of auditory spectral integration. In

future experiments, this can be further investigated by modifying the overall relative intensity of the formant transition in the dichotic condition.

Also of interest are the comparisons between the three sets of virtual F3 transitions in the dichotic condition. The results show that version one was significantly better than either version two or version three at cueing the consonant place distinction. While version one seems to be consistent with previous research, providing evidence for auditory spectral integration and arguing that central mediation is responsible for this integration, the results of versions two and three indicate that there may be more going on than previously recognized. These results suggest that a fusion of the two pairs of sine waves occurs monaurally and is not able to be effectively integrated between both ears. The extent of the auditory system's ability to binaurally integrate separate portions of the virtual signal will be the topic of future investigation in our lab.

While the results of this study provide interesting insight into the mechanisms underlying human speech perception, further studies are necessary to better determine the level of auditory processing responsible for auditory spectral integration. The results of this research may have implications for the development of assistive listening devices and cochlear implants, and are important in furthering our understanding of how the auditory system perceives and identifies the characteristics unique to dynamic speech sounds.

IV. ACKNOWLEDGMENTS

This project was supported by the Ohio State University Colleges of the Arts and Sciences and the College of Social and Behavioral Sciences.

I would like to acknowledge Dr. Robert Fox and Dr. Ewa Jacewicz for their support and guidance, as well as both Chiung-Yun Chang and Erin Saylor, Department of Speech and Hearing Science, for their assistance on this project.

V. REFERENCES

Bedrov, Y.A.; Chistovich, L.A.; and Sheikin, R.L. (1978). Frequency location at the 'center of gravity' of formants as a useful feature in vowel perception. Akust. Zh. 24, 480-486 (Sov. Phys.Acoust., 24, 275-282).

Chistovich, L.A. and Lublinskaja, V.V. (1979). The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research,* 1, 185-195.

Chistovich, L.A.; Sheikin, R.L.; Lublinskaja, V.V. (1979). Centres of gravity' and spectral peaks as the determinants of vowel quality, in *Frontiers of Speech Communication Research,* edited by B. Lindblom ans S. Öhman (Academic Press, London), 55-82.

Delattre, P.; Liberman, A.M.; Cooper, F.S.; Gerstman, L.J. (1952). An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word,* 8, 195-210.

Denes, P.B. and Pinson, E.N. (1993). The speech chain: the physics and biology of spoken language, second ed. New York: W.H. Freeman and Company.

Ferrand, C.T. (2001). Speech science: an integrated approach to theory and clinical practice. Boston: Allyn and Bacon.

Feth, L.L. (1974). Frequency discrimination of complex periodic tones. *Percept. Psychophys.*, 15, 375-378.

Feth, L.L. and O'Malley H. (1977). Two-tone auditory spectral resolution. *J. Acoust. Soc. Am.*, 62, 940-947.

Fox, R.A., Gokcen, J., and Wagner, S. (1997). Neurophysiological and behavioral evidence for a phonetic processor, in *Proceedings from the Panels of the Chicago Linguistic Society's Thirty-third Meeting*, Vol. 33-2, (CLS, Chicago), 311-322.

Fox, R.A., Smith, M., and Jacewicz, E. (2006). Spectral auditory integration and virtual cues to place-of-articulation in stops. *Journal of the Acoustical Society of America*, 119(5), 32-43.

Ladefoged, P. (2001). A Course in Phonetics, fourth ed. Los Angeles: Heinle & Heinle/Thomson Learning.

Lublinskaja, V.V. (1996). The 'center of gravity' effect in dynamics, in *Proceedings of the Workshop on the Auditory Basis of Speech Perception*, W. Ainsworth and S. Greenberg, eds., ESCA, 102-105.

Mann, V.A. and Liberman, A.M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition,* 14, 211-235.

Scharf, B.L. (1972). Critical Bands, in Tobias, J.V., *Foundations of Modern Auditory Theory*, Vol. 1 (Academic Press, New York).

Voelcker, H.B. (1966a). Toward a unified theory of modulation I. Phase-envelope relationship. *Proc. IEEE,* 54, 340-353.

Voelcker, H.B. (1966b). Toward a unified theory of modulation II. Zero manipulation. *Proc. IEEE,* 54, 735-755.

Whalen, D.H. and Liberman, A.M. (1987). Speech perception takes precedence over nonspeech perception. *Science,* 237, 169-171.

Xu, Q.; Jacewicz, E.; Feth, L.L.; Krishnamurthy A.K. (2004). Bandwidth of spectral resolution for two-formant synthetic vowels and two-tone complex signals. *J. Acoust. Soc. Am.,* 115, 1653-1664.

INDEX OF FIGURES